

DOCUMENT RESUME

ED 336 121

IR 053 726

AUTHOR Waters, Donald J.
TITLE From Microfilm to Digital Imagery.
INSTITUTION Yale Univ., New Haven, Conn. Library.
SPONS AGENCY Commission on Preservation and Access, Washington, DC.
PUB DATE Jun 91
NOTE 45p.
AVAILABLE FROM Commission on Preservation and Access, 1785 Massachusetts Ave., N.W., Suite 313, Washington, DC 20036 (\$5.00 prepaid).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Descriptive (141)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Libraries; *Feasibility Studies; Higher Education; *Library Materials; *Microfilm; *Optical Data Disks; *Preservation; Program Descriptions; Records Management; Research Libraries; *Systems Development; Technological Advancement
IDENTIFIERS *Yale University CT

ABSTRACT

The Yale University Library, one of the earliest developers of preservation microfilming practices, is today engaged in the first stages of a long-term and comprehensive effort to develop practical systems for the conversion of microfilm to digital image, and thereby to develop an understanding of the appropriate relationship of the two technologies in future preservation practice. The first section of this report summarizes the results of the initial planning phase of the larger project. In the next section, a vision of the place of digital technology in a research library such as Yale's is articulated. The third section outlines an economy of choices and organization in which the conversion of microfilm to digital imagery might emerge as a viable preservation option. In the fourth section the architecture of a digital imaging system is mapped and critical issues of feasibility are identified. Finally, a plan of work for a demonstration project is presented which investigates the costs and benefits of conversion in the process of building a digitized library of 10,000 books from microfilm copy. (7 tables and figures) (MAB)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED336121

THE COMMISSION ON PRESERVATION AND ACCESS

REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☐ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

From Microfilm to Digital Imagery

On the feasibility of a project to study the means, costs and benefits of converting large quantities of preserved library materials from microfilm to digital images

A report of the Yale University Library to the Commission on
Preservation and Access

by

Donald J. Waters
Head, Systems Office
Yale University Library
New Haven, Connecticut 06520

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Maxine K. Sitts

June 1991

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

1785 Massachusetts Avenue, N.W., Suite 313, Washington, D.C. 20036 • (202) 483-7474

The Commission on Preservation and Access was established in 1986 to foster and support collaboration among libraries and allied organizations in order to ensure the preservation of the published and documentary record in all formats and to provide enhanced access to scholarly information.

The Commission is accessible via BITNET (CPA@GWUVM, SITTS@GWUVM), and by FAX (202) 483-6410.

12053726

COMMISSION PREFACE

In a previous Commission report, Michael Lesk, a member of the Technology Assessment Advisory Committee, asserts that the conversion from microfilm to digital image for preservation of and access to deteriorating library materials is desirable in the long run and is currently technically possible and relatively cheap to accomplish (Image Formats for Preservation and Access, July 1990, page 8). However, document imaging technology is still under development, standards for hardware and software continue to evolve, and the conversion of microfilm to digital image format is largely untested in actual library settings.

It seems critical at this stage that research and investment be directed to the development of systems for managing digital conversion, storage, and access that can be used by university and research libraries. It is to the development of such systems that Yale University Library directed its efforts in this planning study.

The study, conducted under contract to the Commission, explores the feasibility of a project to study the means, costs and benefits of converting large quantities of preserved library materials from microfilm to digital images. The effort identifies requirements for a major, multi-year project to convert materials from film to digital format, to provide both intra- and inter-institutional access to the stored images, and to investigate the broader implications for enhanced intellectual access to digitized scholarly materials. The Commission expects the results to make an important contribution to our capacity to utilize the new digital technologies for preservation purposes. Complimentary copies of this report have been distributed to the Commission's mailing list.

* * *

Published by
The Commission on Preservation and Access
1785 Massachusetts Avenue, NW Suite 313
Washington, DC 20036

June 1991

Additional copies are available for \$5.00 from the above address. Orders must be prepaid, with checks made payable to "The Commission on Preservation and Access." Payment must be in U.S. funds; do not send cash.

This publication has been submitted to the ERIC Clearinghouse on Information Resources.



The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences-Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

COPYRIGHT 1991 by The Commission on Preservation and Access. No part of this publication may be reproduced or transcribed in any form without permission of the publisher. Requests for reproduction for noncommercial purposes, including educational advancement, private study, or research will be granted. Full credit must be given to both the author(s) and The Commission on Preservation and Access.

From Microfilm to Digital Imagery

Table of Contents

Introduction	1
Vision.....	3
The Value of Access.....	3
The Accessible Library	4
• Intellectual and Physical Access	5
• Collection and Collection Renewal	6
Preserving Access through Preservation	7
Digital Imagery.....	8
An Economy of Choices	12
The Focus on Conversion from Microfilm.....	12
A Model of Incremental Investment.....	13
• The Basic System.....	14
• Print on Demand	15
• Online Browsing	16
• Remote Access.....	16
• Browsing at Page Level.....	17
• Browsing at the Document Structure Level	17
System Architecture.....	19
Standards.....	19
System Components.....	20
• Conversion.....	21
• Storage.....	24
• Access.....	25
• Reproduction	26
Feasibility Issues.....	27
Plan of Work.....	30
Project Goals	30
Project Scope	30
Phase 1: Organization.....	31
Phase 2: Setup.....	32
Phase 3: Conversion.....	33
Phase 4: Library Access	35
Phase 5: Remote Access.....	36
Phase 6: Completion.....	37
Conclusion.....	38
Notes.....	39

Tables and Figures

Table 1. Use of Modes of Information Acquisition.....	4
Figure 1. Model of the Access-Oriented Library.....	5
Figure 2. Process of Collection Renewal.....	6
Figure 3. Preservation Microfilming.....	7
Figure 4. Digital Imagery in the Library.....	9
Figure 5. Conversion of Preservation Microfilm to Digital Image ...	13
Figure 6. Imaging Systems Architecture	21
Figure 7. Project Phases	31

From Microfilm to Digital Imagery

Introduction

Over the last three decades, microfilm has become the medium of choice for preserving deteriorating library materials. Microfilm is durable, given a controlled environment. Libraries and archives follow standard procedures and specifications and even share common facilities in the filming process. The costs of the process are predictable, ranging from 10 to 15 cents per page, not including the costs of administrative overhead and of selecting the material to be filmed. Moreover, the technology for gaining access to microfilm is stable and unlikely to change significantly in the future. In short, for the solution it provides to the "brittle-books" problem now plaguing all repositories of material printed on acidic paper, microfilm is essentially risk-free.¹

Michael Lesk, Division Manager of Computer Science Research at Bellcore, in a recent report to the Commission on Preservation and Access, affirmed the use of microfilm for preservation today. However, he took a relatively cautious view of its future role, arguing that microfilm is not likely to be the end of the preservation process. Rather, he argued, it is "a reasonable intermediate step to getting digital imagery," which is a process of scanning books into computer storage.²

Although Lesk acknowledged that digital imaging technology is costly and "has not yet settled down," he argued compellingly that the technology holds a promise that is critical to the library mission of generating, preserving and improving access to recorded knowledge. He observed that conversion from microfilm to digital image is today technically possible and he argued that library readers will, in the long run, demand the results of such conversion. He predicted that digital image technology, including the ability to convert from microfilm, "will be within the reach of most libraries within the decade," and he called for research and investment today in the "development of systems that can be used by ordinary libraries" in the future.³

The Yale University Library, one of the early developers of preservation microfilming practices, is today engaged in the first stages of a long-term and comprehensive effort to develop practical systems for the conversion of microfilm to digital image, and thereby to develop an understanding of the appropriate relationship of the two technologies in future preservation practice. This report summarizes the results of the initial planning phase of the larger project. In the next section, it begins to articulate a vision of the place of digital technology in a research library such as Yale's. Then it outlines an economy of choices and organization in which the conversion of microfilm to digital imagery might emerge as a viable preservation option. The report maps the architecture of a digital imaging system and identifies critical issues of feasibility. Finally, it presents a plan

Introduction (continued)

of work for a demonstration project, which will fully investigate the costs and benefits of conversion in the process of building a digitized library of 10,000 books from microfilm copy.

Vision

For a long view of the future role of research libraries in the university, one must look closely and carefully at the ways that scholars work to create and disseminate knowledge. Several recent studies emphasize that the quality and quantity of scholarly productivity in all disciplines depend critically on ready and timely access to relevant information. The library of the future for such scholars will not necessarily be an "electronic library," nor will its mission be simply to preserve recorded knowledge. The library of the future will aim rather to generate, preserve and improve for its clients ready *access*—both intellectual and physical—to recorded knowledge. Emerging technologies, like digital imagery, will find a critical place there to the extent that they help support the general library mission.

The Value of Access

Beginning in 1985, the Research Libraries Group established a special Program for Research Information Management (PRIMA). The program resulted in a detailed assessment of information needs in over 20 disciplines in three broad groupings: the humanities, the social sciences and the sciences. Among the various needs that emerged from these assessments, many were specific to the disciplines under study. One theme, however, runs through all the studies and applies to all the disciplines. Scholars, in the words of the report on information needs in the humanities, "prefer the familiar and unsystematic methods of consulting footnotes and colleagues."⁴ No evidence is supplied in the reports to support the characterization of these methods as "unsystematic," but it is clear throughout that the value to scholars of consulting footnotes and colleagues is that these methods are readily and easily accessible.

A study sponsored by the Faxon Institute for Advanced Studies in Scholarly and Scientific Communication and presented at a conference in April 1991 makes the same point regarding scientists even more compellingly. The study examined the behavior of a significant sample of scientific professionals as they acquired and used information in their work. It focused on information use by junior, mid-level and senior professionals in the disciplines of chemistry, genetics and computer science in academic, private and government institutions. Relying on diary accounts of so-called information encounters, which are defined as "any use or acquisition of information, ideas, or data that relate to the substance of your profession," the Faxon Institute study contains much valuable and revealing information about the perceived "information competence" of the surveyed professionals, the type of content sources they use to gather information, the frequency with which they use the sources, the time they spend in gathering information, the use to which they put the information, and particularly their methods, or modes, of gathering needed information.⁵

Table 1 summarizes some of the results of the Faxon Institute study and indicates clearly how heavily scholars in the sciences depend for their

Vision (continued)

productive work on sources of information that are readily at hand. The participants in the study found the library to be as useful as consulting one's personal library and as face-to-face discussions. They nevertheless used the library much less frequently than the other modes, presumably because it is relatively time-consuming to get there and use. Moreover, consulting the material at hand in one's own personal library figured in far more "information encounters" than any other single method and was, next to electronic mail, the most frequently used mode. In addition, if all methods of consulting one's colleagues—face-to-face discussions, telephones, written communication, electronic mail and fax—are combined, the Faxon Institute survey suggests that one's personal network of colleagues ranks first and the library a distant third in the percent of times it is utilized to solve an information gathering problem.

Information Mode	Percent of Total Information Encounters in which Mode was used	Average Number of Times used in Past Week	Perceived Usefulness of Mode (average rating on a 10-point scale)
Personal Library/Files	45.0%	9.7	7.9
Library	31.0%	2.8	7.9
Face-to-Face Discussion	29.0%	7.7	7.9
Telephone	15.0%	9.5	7.2
Written Communication	7.0%	4.4	5.9
On-Line Database	6.0%	2.6	6.8
Electronic Mail	5.0%	14.7	6.2
Fax	4.0%	2.6	6.8
CD ROM	2.0%	1.6	5.1
Computer Bulletin Boards	0.3%	3.6	4.3

Table 1. Use of Modes of Information Acquisition⁶

The Accessible Library

The results of the Research Libraries Group and the Faxon Institute studies generally reinforce the findings of almost two decades of research on the information seeking behavior of scholars.⁷ There is still much room for additional work to refine and test these findings rigorously across disciplines and discipline groups. In particular, more detailed information is needed in the context of these studies of the ways that scholars in various disciplines actually use libraries. However, if, as these studies suggest, scholars highly value and tend to favor information that is readily at hand, then a critical measure of success for libraries charting a course into the future is how readily they steer information into the hands of their clients.

A model of the access-oriented library might be depicted schematically as it is in Figure 1. This model is, of course, a simplification. It does not adequately represent all features of the library, and it certainly is not the only way to characterize the system. For the sake of discussion, however, it does highlight the key functional components and their interrelationships.

In this model, collections and services are the two primary components of the library. The library administration is the third major element. The administration supports and undergirds the collection and access services functions by organizing and providing services within the library, such as facility, budget, personnel and technology management, and by representing the library to its clientele and other external agencies. It is also the job of the administration to formulate and articulate the library's mission in the local institutional context.

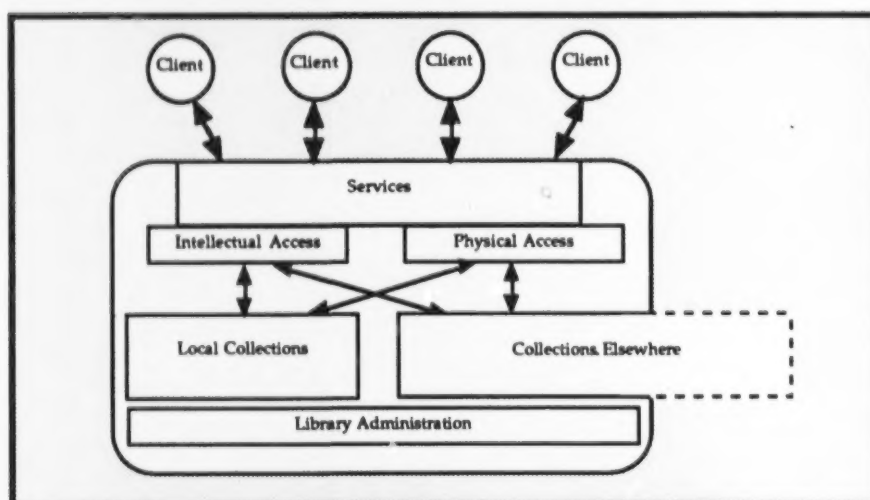


Figure 1. Model of the Access-Oriented Library

The mission of the access-oriented library is to generate, preserve, and improve access to collections of recorded knowledge. This mission guides the fundamental relationship between the access services and the library collections. Access services provide the points of contact between the library and its clientele and are of two kinds.

- **Intellectual and Physical Access**

Intellectual access services, such as cataloging, describe and organize for the library's clients the information and recorded knowledge held in collections locally and elsewhere. The descriptions of a work generally refer in standard ways to such features as author, title, place and date of publication, publisher and format; the organization is typically by subject. Given a subject organization, collocation of materials in the collection by subject makes a range of related material available to library readers and is an intellectual access service that is

Vision (continued)

particularly valued in the humanities and interdisciplinary studies where associative methods make the serendipitous results of browsing the library shelf a critical aspect of research.⁸ Other dimensions of intellectual access include the availability of the content structure of the material, often represented in a table of contents, and the availability of word indices to the full text of the material. Physical access services, such as circulation and interlibrary loan, refer to the methods of distribution that libraries use to put the information and recorded knowledge held in both local and distant collections directly in the hands of their clientele.

- **Collection and Collection Renewal**

The notion of library collections refers to the arrangements and modes of storage of selected information and recorded knowledge. For the library's clients, the collections housed locally are the most relevant. Of course, the collections of any single research library represent only a small fraction of the total reservoir of information and recorded knowledge. Thus, responsibility for the local collections necessarily includes responsibilities both to distinguish them from and relate them to collections elsewhere.

Within the university, the library's collections represent a large, perhaps the largest, capital investment. Like all capital, a library collection is subject to distinct and constant processes of depreciation and renewal (see Figure 2). Material is lost or missing and has to be replaced. New acquisitions expand the collection. The collections are weeded and material is discarded. Existing material deteriorates and needs to be fixed or otherwise conserved, or it is preserved by copying it or by converting it to another format, say from paper to microfilm.

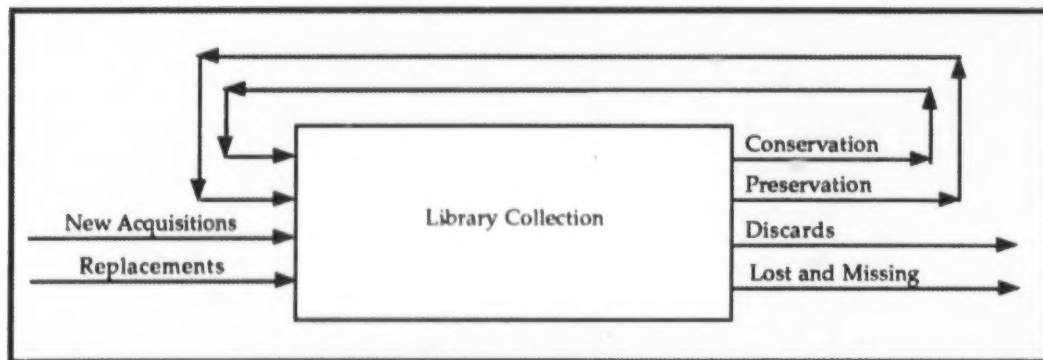


Figure 2. Process of Collection Renewal

In an access-oriented library, each of the various collection renewal processes is subject to rigorous evaluation against the service mission of the organization: How well do they generate, preserve or enhance intellectual and physical access to the collection? For purposes of this discussion, we focus on one of the renewal processes, namely preservation microfilming.

Vision (continued)

Preserving Access through Preservation

Deterioration of a document from age, acidic content, or heavy use limits both physical and intellectual access to it. When a deteriorated item is selected for preservation, it is typically prepared for microfilming (see Figure 3). A photographic process, microfilming faithfully reproduces the original printed material, including stains, discolorations, faded ink, reader notes, and the unsightly borders of overlapping page edges. The process generates a 35 mm film of high contrast, black and white images at a given reduction ratio, depending on the size of the original. The practical limits of using a high contrast, black and white medium mean of course that microfilming cannot adequately reproduce material with color content, shaded gray scale illustrations, or extremely fine printed detail. To avoid loss of information today given current technology, material with such content must be, whenever possible, conserved in its original form.⁹

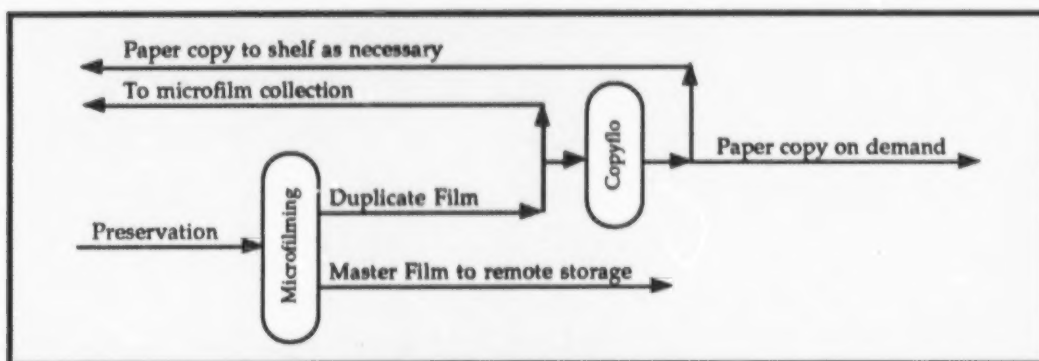


Figure 3. Preservation Microfilming

After filming, a master copy of the film is kept remotely from the library in a controlled environment where, according to some estimates, it will endure for up to 500 years without significant deterioration. A positive and negative copy of the film are stored in the microfilm collection of the library. The film can be further duplicated from the negative copy with relative ease at a cost of approximately \$20 per reel and the Copyflo process can generate a paper copy at a cost of approximately \$0.15 per page. For high use items, the library may decide to return a paper copy to the library shelves. Library readers may obtain a full paper copy on demand, or they can print selected pages at a microfilm reader/printer.

Faced with the total loss of intellectual content in a deteriorating document, the library realizes distinct advantages from preservation microfilming. The content is saved for the collection in a durable and compact form using a technology that is relatively simple and well established. However, readers lose a significant measure of intellectual and physical access to the material in the bargain.

Vision (continued)

Unlike a book, which one can carry away and use virtually anywhere, microfilm confines the reader to use special projection equipment at a specific location. A reader can learn of the intellectual content of a document in microform, in part, through the bibliographic entry created for it during the preservation process but, by converting paper to film, the process splits a collection apart so that one cannot as readily benefit intellectually from the physical association of a volume on the shelf with others of related content. Moreover, microfilm is cumbersome to browse and difficult to read. With a microfilm reader, one cannot as readily as with a paper copy use the internal structural apparatus of the volume—its table of contents, chapter summaries, prefaces, footnotes, indices, etc.—to gain rapid and efficient access to the intellectual content of volume, be it a fact, page, chapter or some combination of these.¹⁰ Finally, printed copy from microfilm sources using either the Copyflo or reader/printer technologies has won little praise for its qualities of clarity and resolution.

Not surprisingly, given these deficiencies, readers complain intensely about using microfilm. In one sense, these complaints about the deficiencies of microfilm seem unfair: without preservation microfilming, substantial portions of library collections would be lost entirely. The inconvenience of using microfilm in this context seems a small price to pay. However, given the value to scholars of information accessibility, the library is obliged to insure that it can apply no better and cost-effective mix of technology than microfilming to save the intellectual content of rapidly deteriorating materials and, at the same time, preserve or improve scholarly access to the material. For this reason, libraries like Yale's, in which over 80% of the collection is currently or potentially brittle, must investigate the means, costs and benefits of adding to its preservation toolkit new and promising tools, including the technology of digital imagery.

Digital Imagery

Digital image technology provides the means of digitally encoding scanned documents in image form for computer-based storage, transmission and retrieval. The digital images produced using the technology contain text, but the text is not converted to and is not accessible in alphanumeric form. The potential exists for such conversion, in whole or in part, by applying character recognition technologies to the digital images. The use of digital imagery envisioned here is designed to anticipate the eventual application of character recognition but it does not presently incorporate it.¹¹

Digitization holds the promise of greatly improving the accessibility of preservation materials compared to microfilming (see Figure 4). Physical access to the library of digital images, like access to microfilm, requires special equipment. Compared to microfilm readers, however, digital image equipment—a powerful computer workstation with a relatively high resolution monitor—is general purpose technology for which access to the digital library might be only one application. Moreover, the scholar who is outside and possibly even

Vision (continued)

off the campus of the library holding the images can gain access to them relatively easily and quickly because the digital images can be transmitted over the high-speed networks that many university campuses have already installed and interconnected. The scholar also has the opportunity, using digital image technology, to add all or part of the image document to his or her own personal library by requesting a printed copy. The printed form is likely to be of better quality and higher resolution than similar output from microfilm, and it may even be substantially cheaper, again because the underlying technology serves a general purpose.

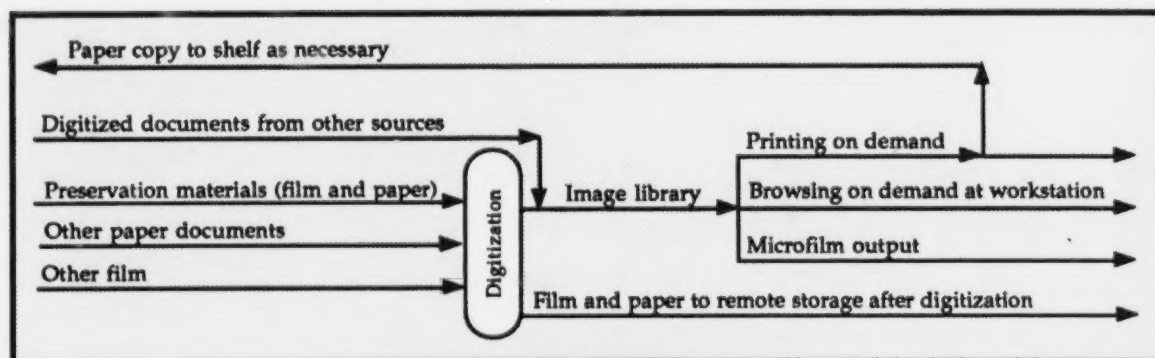


Figure 4. Digital Imagery in the Library

Digital imagery also promises to substantially improve intellectual access, compared to microfilm. The ability of the technology to put a high quality, relatively inexpensive paper copy of the original document into the personal library of scholars on demand makes the standard intellectual apparatus—the table of contents, index and so on—directly accessible. Moreover, unlike the current practice with images on 35 mm microfilm, digital images can be indexed by page numbers and, if resources permit, by higher order document structures such as section, chapter, and part. Given such an index and the software to use it, scholars can move quickly through the digital document and browse its intellectual contents much more efficiently than they can with documents preserved on microfilm.

In the forms outlined here, digital imagery affords no substantial improvement over microfilm in providing an effective substitute for shelf browsing. However, recent studies suggest that by adding supplementary information, such as transcriptions of tables of contents, to online bibliographic descriptions, one may be able to provide even more fruitful opportunities for intellectual associations and serendipity than is possible from the traditional methods of subject classification and collocation of paper material by subject on library shelves.¹² As content material is added to online bibliographic records and the resulting benefits of improved intellectual access to materials in all formats become apparent, it may be possible to achieve substantial

intellectual leverage from material in digital image format by creating direct electronic links between the supplementary information in the bibliographic record and the digital images of the document to which it refers. The creation of such links will require considerable work, particularly in the online bibliographic system, but the vision of their possibility means also that one must carefully and creatively design the page number and document structure indices within the digital image system so that the digital images can be easily and reliably related to the content information stored in other machine-readable forms, such as the online bibliographic record.

Our vision of the library of the future hinges on the central distinctions and interrelations that we have posited among

- the library collections, which comprise the modes of storage for selected information and recorded knowledge;
- physical access, which comprises the modes of collection distribution; and
- intellectual access, which comprises the principal modes of collection use.

Just as we expect digital image technology to introduce more flexibility into and otherwise enhance the modes of collection use and distribution, particularly compared to microfilm, we also expect the technology to exert similar effects on the modes of storage in the collection. In this context, it is critical to distinguish between archival storage and storage for use and distribution.

In the library of the future, we presume that microfilm will continue to play an important role, perhaps still as the preferred medium for storing an archival version of preserved documents.¹³ For the use and distribution of preserved materials, however, digital images will likely become the preferred storage medium. Depending on costs, microfilming may remain a preliminary step in a preservation process leading eventually to digitization, or film may be generated from the digital image. In either case, we expect the digital technology to accept as input on demand the substantial base of documents already preserved on microfilm and now available only in that form.

Finally, our vision of the use of digital image technology suggests several wider effects. Although we expect to establish the digital library out of a primary impulse to preserve and improve access to the huge quantities of deteriorating and brittle books on library shelves, we also expect the library to be enriched over time with documents in digital image form that have been created for various other reasons both in and out of the library. In addition, digital image technology will likely have a significant, perhaps profound, effect on the ways in which the library of the future views, organizes and values its spaces and staffs. For example, to the extent that the digital library arises as a way of renewing and replacing large stocks of paper and microfilm materials, and to the extent that the digital library is stored and accessible remotely from the library proper, the use of the central stack

Vision (continued)

space of the library will almost certainly change and the balance of staff functions traditionally associated with circulating and shelving collections stored in paper will likely shift to modes of delivering documents generated on demand from electronic formats.¹⁴

An Economy of Choices

Libraries of the future will not attain in one quick step the vision articulated here of the use of digital technology in preserving and improving access to recorded knowledge. The promise of service benefits and possible cost savings will move libraries toward the technology in various ways. Budgets and other organizational impediments will constrain them from investing in it in still other ways. If the promise justifies the investment, most will incorporate digital imaging technology just as they have adopted and assimilated other technologies in the past: through a series of particular and incremental decisions and choices tailored to the mandate and needs of their specific institutions.

Those who develop digital image technology for use in library preservation need to be mindful of the principle of incrementalism and of its power and influence in organizational change. The economy for managing and administering library resources, like that of most large institutions, is an economy of incremental choices, even about potentially revolutionary technology. Development work in the application of digital imagery thus needs to yield results about the means, costs and benefits of digital imagery that will facilitate and inform, not distort and mislead, the operation of such an economy.

This section attempts to identify a critical set of functional choices, which will likely govern the incorporation of digital imagery in library preservation processes. The next section outlines many of the detailed technical requirements for a system architecture that can flexibly accommodate the types of choices libraries will likely make. The final section lays out a plan of work to develop the architecture and to generate the information that libraries will need to make prudent choices about the costs and benefits of image technology.

The Focus on Conversion from Microfilm

Cornell University, an early pioneer in the use of digital imagery for library preservation, chose to investigate ways of overcoming the limitations of preservation microfilming by scanning deteriorating documents directly into image form, rather than microfilming them. Additional work still needs to be done to expand the means of direct input, including methods for handling oversize, colored and other kinds of special documents. Moreover, because they can be enhanced electronically to eliminate stains and other markings and to heighten the contrast of fading ink, digital images might ultimately prove a better quality source for microfilm copy than the original paper source. The means of generating microfilm output from digital image thus also needs to be explored systematically.

Paper, however, is not the only possible source of input in the creation of a digital library of preserved materials. The Yale University Library seeks to demonstrate for the wider library community the viability of one alternate source by developing the means and identifying the costs and benefits of scanning preserved materials from microfilm sources. Although it will take input from microfilm rather

An Economy of Choices (continued)

than paper, the imaging system developed in the Yale project will otherwise provide the same basic functional capabilities as the one being developed at Cornell. In the fully developed system, library readers will be able to browse the image library from a computer workstation (see Figure 5). They will also be able to obtain a printed version of the digitized document on demand; the library may itself generate a paper copy of the preserved document to return to the library shelf. Moreover, the image library will be stored remotely from the shelf collection. Because of the presumed accessibility of the digital library, both in and out of the library, the microfilm version of the digitized documents can ideally be moved out of the library to more remote and cheaper storage.

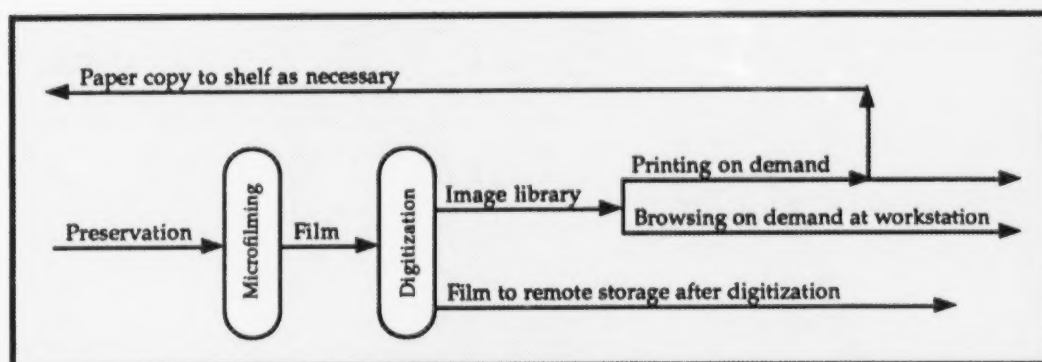


Figure 5. Conversion of Preservation Microfilm to Digital Image

The system at Cornell is being developed initially to handle an image library of 1,000 volumes. In addition to developing an alternate source of input, the Yale project will test the scalability of imaging systems like that at Cornell by increasing the digital library an order of magnitude to 10,000 volumes. The size of the digital library at Yale, compared to that at Cornell, will help demonstrate the economies of scale in the system and help insure for the library community that realistic measures are made of the costs both of adding components incrementally to the system and of maintaining them over time.

A Model of Incremental Investment

The costs and benefits of an imaging system depend largely, though not exclusively, on the major functional components of the system and on the method and timing of their implementation. One does not have to adopt all possible components to create a working system. Given the mission the system is intended to serve, some components are essential, and one must implement them; others are less critical, and one can install them optionally or later on an incremental basis. The following analysis distinguishes the major functional components of a preservation imaging system, orders the components in relation to the purpose of the system, and arrays them in a cumulative sequence of steps. The sequence suggests a plausible strategy offering choice at each level for a library investing in imaging technology. To support and

inform the choices, the development of the imaging project at Yale must be designed to measure the incremental costs and benefits of the system components identified at each stage in the sequence.

• The Basic System

If the aim of preservation is to preserve and improve physical and intellectual access to deteriorating printed material, and assuming that the material has already been microfilmed, then the simplest and most straightforward imaging system is one that produces from microfilm a high quality printed copy of the original document for the library to return to its shelves. The essential components of such a system would be:

- a microfilm scanner;
- a temporary storage device to hold the digitized images;
- a quality control station to review the images, enhance them if possible and rescan the film if necessary;
- a high speed network for transferring the images to the print service;
- a print service, for queuing and otherwise controlling the print job; and
- a high quality digital printer.

A service bureau could, of course, provide all or part of these functions. It could scan with quality control and print the document; it could scan with quality control and return the set of digitized images to the library for printing; or it could accept the digitized images from the library for printing. It will be important for the Yale demonstration project to gain sufficient in-house experience with digitization so that one can fully convey to a service bureau the standards and expectations for conversion and indexing and so that one can fully compare the costs of an in-house process with the costs of hiring a service bureau to perform the same functions.

Because the Copyflo process is already capable of producing printed copy directly from microfilm, a library will likely find this basic imaging system attractive if and only if the costs of digitization and printing are less than the costs of Copyflo, or the output from the digital printer is of higher quality. Assuming, as is likely, that the resolution of the digital image generated from microfilm is in the range of 300 dots per inch (dpi), then the print quality will surely match or beat Copyflo quality. Michael Lesk has estimated, moreover, that the cost of a simple conversion from microfilm to digital image is approximately \$0.02 per frame.¹⁵ Early estimates from Cornell suggest that the costs of printing a document from image form will be approximately \$10.00 per 300-page book (unbound) or \$0.033 per page.

These estimates carry with them assumptions about labor costs, productivity, volume and methods of equipment financing and amortization that may or may not hold in all circumstances. However, recognizing the limitations of the estimates, compare them still to the

present cost of \$0.15 per page (unbound) for the Copyflo process. An imaging system stripped to its essentials and capable simply of digitizing and printing microfilm images not only will produce higher quality output but also may yield more than 60% in savings over the current means of printing from microfilm. Of course, with added components, an imaging system can deliver much greater function than the basic system described here, but the stakes of the investment rise correspondingly.

• Print on Demand

If a basic imaging system gives the library the ability to preserve access to a deteriorating document by generating a high quality, relatively low cost facsimile copy to return to the library shelf, then one may legitimately wonder what additional investment it would take to produce the printed copy on demand for a scholar to put in his or her own personal library. An imaging system capable of on-demand printing would need to add the following components to the basic system:

- a permanent storage mechanism for the document images in digital form;
- a title level index to the document in image form, so that one can distinguish one document in the storage medium from another;
- an entry in a new or existing bibliographic statement, presumably recorded in the library's online catalog, which indicates to the reader that the document is stored in digital form and is available for printing on-demand;
- a print request service; and
- an image service that extracts the images under the unique identifier and transmits them to the printer.

For the incremental benefit of on-demand printing, one also incurs substantial costs, particularly in the storage mechanism. The storage of digital images is technically complex and will be treated in greater detail in the system architecture discussion in the next section. It is important to note here, however, that the notion of "permanence" in digital storage implies the cost of periodically renewing the image files—and relatively more frequently than books and microfilm—as storage media, file formats and the equipment used to gain access to the files all change. It is possible, but not yet proven, that savings in space from denser and more compact storage and other efficiencies gained from the changing technology could balance the costs of renewing image files and thus make the renewal process self-sustaining.

The indexing system at this stage is simple to create and use. It consists of applying a classification number or other unique identifier to the collection of images that comprise the digital document. The unique identifier provides the key that links the bibliographic entry in the catalog to the title in the digital library, much like a call number enables a reader to go from the catalog to a book on the shelf or to a microfilm roll in a specific cabinet drawer.

A scholar cannot, of course, always tell from the catalog entry whether a document is relevant to the problem or research at hand. Before requesting a printed, personal copy, one still may need to browse the microfilm document in the library to determine its relevance. Investment in an imaging system that provides a print-on-demand service thus incurs the costs of storing and indexing the documents in digital form while still requiring the library to maintain the microfilm version of the document for the reader to browse.

- **Online Browsing**

Given permanent storage for a print-on-demand service, another natural extension of the imaging system would be to enable the reader to browse a document online directly in digital image form. The capacity for online browsing requires, in addition to those already identified, the following components:

- a computer workstation and printer at which a reader can retrieve, display, move back and forth through, and print selected images from the document; and
- a presentation service that tailors the image presentation to the particular type of browsing workstation being used.

The incremental investment in these components of the imaging system affords the reader functional capabilities that, in at least one respect, resemble browsing a document in microfilm form. Because the image document at this stage is indexed only at the title level with a unique identifier, one cannot go automatically to a specific page or section of the document. As with a microfilm reader, one must, in effect, crank back and forth through the document image by image. The major functional difference from microfilm use at this stage is that the reader does not have to find and load the film manually, but rather depends on the system software to find and load the document automatically. In addition, after browsing, the reader can request that the system generate a high-quality, printed copy of the document for personal use.

It is incumbent on the system developers to insure that the browsing and print-on-demand features incorporated in the imaging system at this stage of investment are compelling and sufficient for the scholar to use instead of microfilm. Assuming that they are, the library can begin to accrue at least modest savings in space and service costs by eliminating the on-site microfilm copy of the document. The library can also contemplate significant extensions of the browsing capability.

- **Remote Access**

The library could extend the browsing capability of the image system in one way by making it accessible remotely from the library. Within the library, presentation of the browsing functions can be controlled by limiting the type of workstation used. Remote access, however, would require an enhanced presentation server to support the variety of image-capable workstations that scholars use outside of the library. The image server and the local campus network would also need to be able to bear the increased traffic. Whether the browsing function can practically be extended beyond the campus network is a subject of

An Economy of Choices (continued)

considerable interest that will require additional research and much testing.

- **Browsing at Page Level**

The library can also extend the browsing capability of the image system by making the images of a document directly accessible by page number. The incremental investment in page level browsing depends directly on the costs of generating, maintaining and supporting in the browsing software a page level index to the image document. One cannot issue a command to go to a specific page in a document until and unless someone deliberately takes the time to create an index linking an image with a document page number. The pages may be numbered in roman or arabic notation. On the other hand, some pages may not even be associated with a specific number in any form of notation. The indexing and retrieval system needs to reflect nevertheless these distinctions, variations and peculiarities.

- **Browsing at the Document Structure Level**

The library can extend the browsing capability of the image system even more deeply by making the images of a document directly accessible by its own internal structural divisions, such as title page, table of contents, part, chapter, section, and index. That is, without moving page-by-page through the document, the reader could issue a command to go immediately and automatically to the third chapter, to part three, or to the title page. Again, the incremental choice to implement an imaging system with this level of functionality depends directly on the library's interest in creating an index linking an image or set of images to a particular structural element or elements of the document. A slight variation on the theme would provide the reader the capability of temporarily structuring the document during the browsing process with placeholders or bookmarks. An even more ambitious variation would enable the reader to structure the document completely or partially in his or her own terms and save the resulting structural index under his or her own name for later use.

The analysis presented here of a library imaging system that is intended to preserve and improve access to deteriorating materials and is derived from the conversion of microfilm to image format does not exhaust all possible uses of such a system. For example, one might elect to digitize an entire collection of material on the grounds that the intellectual coherence of the material justifies the effort to make it more accessible. Alternatively, to help mitigate the costs of storage and indexing, one might opt to digitize only heavily used materials, as an intermediate step, or even as an alternative to putting a paper copy back on the shelf. Or one might opt only to digitize materials as they are requested for use, leaving the unused portions of the collection in microform.

Nor does this analysis exhaust all possible features of a preservation imaging system. For example, once character recognition technology sufficiently matures, additional enhancements to the imaging system

will become possible. All or part of a document in image form may be converted to a text file of alphanumeric characters and the document images could potentially be linked to a portion of the text file, such as the table of contents, which is stored in the bibliographic description of the image document, or to a keyword index generated from the entire text file. Advances of these kinds could greatly enrich the intellectual value of the image library.

The analysis provided in this section does, however, highlight the major functional dimensions of a preservation imaging system. It also suggests a plausible array of incremental choices that might motivate a library to invest economically in such a system. Of course, one must distinguish motivation and functional need from both the underlying technical architecture of the imaging system, which serves the library purposes and meets its requirements, and from the plan of work for a demonstration project, which is designed to develop a fully articulated system within the architecture and to generate sufficient information about the system so that libraries can make prudent, practical decisions about its costs and benefits.

System Architecture

The architecture—the overall design and structure—of a system for creating, storing, retrieving and printing library documents in image form has evolved gradually over the last decade. Early research and the development of pilot projects, particularly at the Library of Congress and the National Library of Medicine, have tested, refined and validated key design ideas as the imaging technology has emerged: workstations perform imaging functions; they are distributed over a high-speed network; and they use needed resources on the network and interact with each other in a client/server fashion.¹⁶ In its digital preservation project, Cornell has incorporated a distributed workstation, client/server design in its system architecture, and Yale will also.

Other, more general design principles also apply to the architecture of imaging systems. Several of these principles will govern the system being developed at Yale, and one needs to articulate, rather than assume, them. First, the data, both the images and the indices to them, will comprise the primary asset generated in the Yale project. The software and hardware needed to create, store and use the data will change and be replaced. The data will remain, however, and must be readily and easily convertible from one system to another. Second, building on the argument in the previous section about the need for choice, the system components must be developed in modular form so that they can be readily and easily added, eliminated or changed as needed. Third, to insure flexibility when adding to, eliminating from or changing the imaging system at Yale, the components used must conform to official or de facto industry standards, or be built to standard interfaces.

Invoking these principles in reverse order, this section identifies the relevant standards for the imaging system at Yale, reviews the system components and the technical issues associated with them, and addresses the overall feasibility of the design by highlighting areas where concern for the data might warrant special attention to specific system features.

Standards

The relevant standards that apply to the design and construction of the digital imaging system at Yale fall into three groups: communications, storage and application standards.¹⁷ As most major research universities in the United States have done, Yale University supports a variety of networking protocols, but has targeted the so-called ethernet standards, including TCP/IP, as the protocol suite of choice for network communications. Components of the system may operate in a sub-network of the campus network using a different standard protocol suite, for example token ring at the data link level. However, the sub-network must supply a reliable and well-behaved gateway to the central ethernet-based network spine for communication with other components of the system.

System Architecture (continued)

For the storage of image data, the system at Yale will support the Tagged Image File Format (TIFF), which provides the header level standardization that is needed to allow for the exchange of images. Because image files are so large, a standard is also needed for compressing the files to some fraction of their original size in order to store and transmit them economically. The CCITT Group 4 compression standard is deficient in some respects because it does not cover pages that are larger than letter-size, nor does it handle well images with extensive gray scale or color. Standards groups are currently cooperating to develop more adequate compression standards. In the meantime, the CCITT Group 4 standard should serve the needs of this project.

At the applications level, bibliographic information referring to a document in image form will be entered in the library's online catalog in standard machine-readable cataloging (MARC) format. Existing MARC conventions are not completely adequate for describing a digital image version of document, but the appropriate standards organization has drafted a suitable revision of the standard. Yale will adopt a temporary solution until the revised standard is accepted.

Two other standards are also appropriate at the applications level. First, indices that open an image document to browsing at the title level, at the page level, and at the document structure level depend on work that describes the contents of the images and defines and makes it possible to exchange and present a particular set of images as a set of higher order entities (e.g. a book, a chapter, a set of pages). Standards covering document content need to apply (or be appropriately extended so that they apply) to the creation, storage and use of the various levels of structural indices. Several vendors, including IBM, have promulgated their own internal document content standards. However, ISO 8613, entitled "Information Processing—Text and Office Systems—Office Document Architecture (ODA) and Interchange Format (ODIF)," is the prevailing international standard, and the Yale project will seek to conform to it.

Second, the structure files, which describe the content of image documents in conformance with ISO 8613, need to be stored in and retrieved from a database file linked referentially to the files of document images. We would expect the database to be relational. We would also expect it to incorporate and conform to the Structured Query Language (SQL) standard.¹⁸

System Components

A general representation of the system architecture for the preservation imaging project at Yale appears in Figure 6. The components all are linked to the campus high-speed ethernet network, which has a fiber-optic core and current transmission speeds of up to 10 megabits per second. The key components of the design are a conversion subsystem, a reproduction (or printing) subsystem, a storage subsystem, and an access subsystem.

System Architecture (continued)

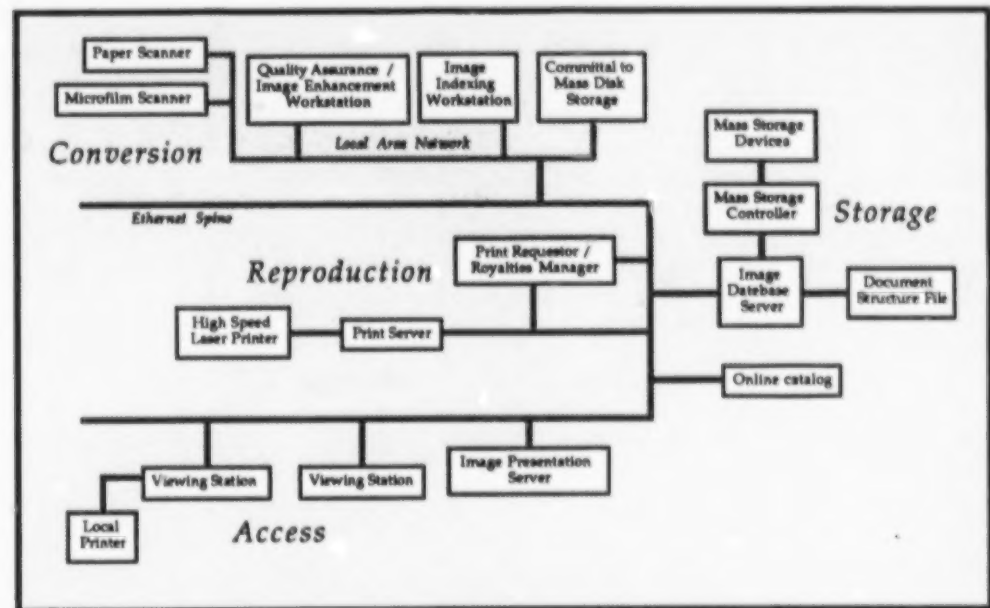


Figure 6. Imaging System Architecture

- Conversion

The purpose of the conversion subsystem is to accept as input a series of microfilm frame images that constitute a document (book, pamphlet, or periodical), and to take a "digital photograph" of each frame so that the image is stored as a computer file and can be displayed on a computer monitor. The conversion process assures the physical quality of the digitization, which includes controlling resolution as well as image content, contrast, skew and sequence. The process also serves to characterize, or index, the set of image files so that they retain their identity as a single document for subsequent storage and retrieval. The final step in the process is to commit the set of files as a document entity to permanent storage.

The conversion subsystem includes a microfilm and paper digitizer and, at minimum, a controlling workstation at which an operator can perform the quality control, indexing and commitment functions. To provide maximum flexibility, the workstation and digitizer should be linked via a high speed local network. With the network interconnection, additional workstations could be added to divide the labor, but each function could still be performed at every station. Alternatively—and this may be necessary for maximum efficiency—the digitizer may be directly linked to a controlling workstation, at which an operator performs the the quality control functions. Upon completion, the operator could then hand off the files, using a network (or "sneakernet") file-sharing mechanism, to another workstation for further conversion processing.

The digitizer must accept negative or positive 35 mm roll film in comic or cine mode at various reduction ratios. It should provide an automatic feed mechanism allowing unattended operation. It should be equipped with sensors to detect the beginning and trailing edge of each frame and to correct automatically for framing skew. If possible, it should detect the beginning and trailing edge of each page on frames filmed with two pages per frame. For the effort expended in digitizing, the Yale project will seek output at the highest resolution that is technically possible. The digitizer should nevertheless be capable of scanning at different resolutions and, to prevent labor bottlenecks, it must be capable of generating its highest resolution output at a rate per frame of 2 seconds or less. Finally, the digitizer should present an open set of interfaces that would allow it to work with DOS, UNIX or Macintosh workstations. A high-speed paper digitizer should be on hand with a similar set of features to allow for controlled comparisons of speed and quality in cases where the paper copy of the microfilmed material is still available.

To work with the digitized images, each workstation in the conversion process needs the processing ability to read and expand a compressed page image and to compress and save a displayed page image to disk. All the workstations need to be able to move backward and forward through the sequence of image files and to jump to a specific image relative to the current image. Each workstation also needs a high resolution monitor capable of displaying a readable page image. Each workstation should provide general image manipulation functions such as zoom, shrink, pan, scroll, and rotate. Finally, all workstations in the conversion process need access to a local laser printer for printed copy of selected pages.

The workstation controlling the digitizer and the quality of its output should enable the operator both to initiate a completely automatic scanning process and to intervene in that process cleanly and robustly. It should afford the operator the ability to adjust and set the scanning resolution, perhaps producing one set of images at a high resolution for printing and another set at lower resolution for browsing. The workstation should support techniques for automatic image enhancement, including fixed and dynamic thresholding and peak detection, which help to improve poor print-to-background contrast. In addition, it should permit the operator to crop image borders and perform other real-time image enhancement functions, as well as to rescan an image completely as necessary. Finally, the workstation should automatically create files from the digitizer in TIFF, compress the scanned images by the CCITT Group 4 two-dimensional technique and assign sequential names for the image files.

The indexing application in the conversion process should guide the operator through a process of providing information about certain key aspects of the structure and content of the set of images comprising the converted document. In one window, it should display the image set

and enable the operator to move backward and forward through the sequence of image files and to go directly to any specific image. In another window, the indexing application should prompt the operator to verify that the sequence of images is correct and to supply a unique document identifier under which the set of images can be stored as a single entity. When supplying the identifier, the application should check for uniqueness against the online list of existing identifiers.

In a third window, the application should give the operator access to the local online catalog, create an appropriate entry for the document in its digital form, and indicate the unique identifier by which the digital document may be found. When the bibliographic work is complete, the operator may optionally supply information that relates the image sequence to a page number notation and sequence. He or she may also index the significant structural divisions within the document. The operator then should be able to save the structure and content information in a database file, which is distinct from, but points to, the set of images that constitute the preserved document. If two copies of a document exist in different resolutions, the structure file should accurately point to the copies as different instances of the same document with the same structure.

Once the structure file is created, and even after it is permanently stored, the indexing operator should be able to recall and modify it. Moreover, once the operator has created a page number index for a specific document, he or she should be able immediately to issue a command to go to a specific page. Similarly, once he or she has created an index to the internal structure of the document, the application should make it possible immediately for the operator to request to go to the image at the beginning of each structural division. The interface should provide the operator with a special "bookmark" function, which would mark a specific image and allow a direct return to it at any time; the location of operator bookmarks, however, should not be saved in the permanent version of the structure file.

In applying the committal function, an operator verifies that the conversion process is complete and that the resultant set of digital images may now be stored. The function assumes, as a matter of security and user convenience, that the image documents and the associated structure file database are stored in a central, network-accessible, storage facility; not, by analogy to microfilm, in cabinets near the browsing workstations. In the event that the purpose of the conversion is not to create a permanent copy of the document in digital form but to print it, the committal function may close the process by transferring the images by removable media or over the network to the high-speed digital printer.

With respect to disk storage in the conversion process, each workstation that controls the digitizer needs sufficient working magnetic disk space for the operator to store and manipulate the uncompressed, scanned images of at least two books (approximately 600 megabytes). At each

System Architecture (continued)

workstation that performs the indexing and committal functions, there needs to be magnetic disk space to store the compressed images of at least two books (approximately 100 megabytes). At workstations performing the committal function, there should also be a portable storage medium, such as an optical WORM (write once, read many) drive holding a disk platter, to which the compressed document images can be written and then physically moved to the permanent disk storage site. Alternatively, there needs to be the capability to write to a mass storage facility via the high-speed network.

- Storage

The purpose of the storage subsystem is to provide a high capacity, reliable and responsive medium (or mix of media) in which to store a collection of documents that have been preserved in digital form. A conversion operator must be able to transfer a digital document to the storage subsystem, and from it one must be able to retrieve documents either to print or to browse at a computer workstation. The components of the storage subsystem include the mass storage devices, the mass storage controller, the document structure file and the image database server.

Assume that the conversion process creates each image file at a resolution of 400 dpi and that in compressed form each image occupies 150 kilobytes of storage. Assume also that each book that is converted consists of 300 pages. It follows by simple multiplication from these assumptions that a digital library for the 10,000 books slated for conversion in the Yale project will require approximately 450 gigabytes of mass storage.

One assumes that optical WORM is today the most cost effective medium for storing this volume of information.¹⁹ However, if a second copy of each document is created at a lower resolution for browsing purposes, the storage requirements may be substantially less for the second copy (about 125 GB) than for the first, and use may be frequent enough to justify the costs of storing the image files on magnetic disk, which boasts of faster access times than the optical medium. The WORM cartridges holding the higher resolution versions for printing may be mounted in a jukebox, or held off line requiring an operator to intervene to mount them when requested. Data on magnetic storage would have to be regularly backed up to tape as a precaution against data loss from a disk failure; data on WORM media would have to be periodically copied to new media as the formats or access equipment become obsolete and are replaced.

The mass storage controller in the subsystem provides an interface between the disk storage and the image server. It translates requests for images into commands which locate and retrieve document images from the magnetic storage device or the optical jukebox, as appropriate. In the event that the document is stored offline the controller will generate a request to the operator to mount the appropriate cartridge. For requests to store image documents, rather than to retrieve them, the

controller will locate space, appropriately store the files and update the file directory.

The document structure file contains information indexing aspects of the structure and content of each image document. It is an integral component of the image library and information from it must accompany each request for a document; conversely, any image document being newly stored must bear an index component for storage in the structure file. The file is a relational database conforming to the SQL standard. As an online file, it requires regular backup.

The image database server links the document structure file through the storage controller to the document image files. It is network accessible and responds to requests for images by retrieving the relevant information from the structure file and the relevant files from the image library and then spooling them for delivery to the access subsystem or the printer as appropriate. The server also responds to requests to store image documents by insuring that the structure file is updated and that the image files are spooled and transferred to the appropriate mass storage device. The server needs a substantial amount of memory to cache the inbound and outbound documents. A critical function of the server in this process is to deliver status data to its clients, including information regarding transmission progress, and errors in storage or retrieval.

- Access

The purpose of the access sub-system is to provide a facility for individuals to retrieve and use the library of documents preserved in image form. The components of the subsystem include the image presentation server and the image viewer stations. In the simplest case of the subsystem, image documents need to be presented for viewing at a single type of workstation which the library provides and the selection of which it controls. The more complicated case results when the library opens access to the image documents to scholars across the campus who use different kinds of workstations with varying image display capabilities. In either case, the minimal functions provided by a viewing station should be the same. The station should enable the reader to retrieve an image document by its unique identifier and to display and browse the document depending on the way that it is described in the structure file.

In order to be "image-capable," each viewing station needs the processing ability to read and expand for display a compressed page image. The stations require a high resolution monitor capable of displaying a readable page image. Each workstation should provide general image manipulation functions such as zoom, shrink, pan, scroll, and rotate. Finally, viewing stations should be able, if the equipment is available to them, to transfer images to local laser printers for printed copy of selected pages.

From the access workstation, the reader must be able to search the online catalog to determine a document's unique identifier. Given the

System Architecture (continued)

identifier, he or she then should be able to request the document from the image server and have it available for viewing on the screen. In the absence of any additional index information, the reader should be able to move backward and forward through the sequence of image files and to go directly to any specific image relative to the current one.

However, if a page number index exists, he or she should be able to issue a command to go to a specific page. Similarly, if an index exists to the table of contents, the parts and chapters of the document and other major parts of its internal structure, the software should afford immediate access to the image at the beginning of each structural division. The interface should also provide the reader with a special bookmark function, which would mark a specific image and allow an immediate and direct return to it at any time; the location of reader bookmarks, however, would exist only for the duration of the browsing session.

The image presentation server separates these various access functions, which all viewing stations must provide in common, from the different ways that the operating systems of various workstation platforms (DOS, UNIX, Macintosh) enable them to be presented effectively to the reader. At the beginning of a browsing session, the server will interrogate the particular viewing station to determine its type and capabilities. It may be that the station is not adequately equipped for viewing images and the server will respond accordingly. Otherwise, when the image server responds to a request for a document, the presentation server will spring to life. It will take the document structure file, which describes critical aspects of the document and, using the conventions of the Office Document Architecture, will provide a high-level set of directions for how document content should be presented. The server translates these directions into a format and set of program instructions that are suitable for the particular workstation. Using a driver it obtains from the server, the viewing station, in turn, processes the programmed instructions and generates a presentation format in which the reader can easily browse the image document.

- **Reproduction**

The purpose of the reproduction subsystem is to generate very high quality printed copy of all or part of selected documents that are stored in digital image form. Components of the subsystem include the printer, a print server and a print requester. The printer must be able to take as input image files with very high resolutions (600 dots per inch or better). It must generate printed output at the same resolution. For greatest economy, it should operate at very high speeds (100 pages per minute or better is a current benchmark criterion) and provide built-in collation and binding capabilities.

The print server manages the queue for the printer, providing various operator controls including the ability to set job priority, inquire about job status, and cancel jobs. It should provide a substantial spooling capability, and it must accept TIFF files compressed according to the CCITT Group 4 standard. It may need to reformat such files to an

System Architecture (continued)

internal format, but should do so with no loss of quality or information. To accommodate print requests directly from the conversion subsystem, the print server may need to have a compatible WORM drive for direct input of the document to be printed.

Users who want to print all or part of a particular document will typically do so by invoking a network service to submit print requests. An individual can request a printed copy of an image document by the unique identifier specified in the online catalog. The request service will retrieve the structure information regarding the document and ask the individual to determine what portions or what pages to print and whether binding is required. The print request service should have built-in accounting capabilities so that individuals can be billed for the copies that they generate. General use of the request service on the network for billing presumes that there is a way to authenticate users and verify that they are who they say they are. Assuming the authentication issue is addressed, a natural extension of the print request service would be to use it to collect appropriate charges for the reproduction of royalty bearing documents, such as recent journal articles, which may eventually be added to the digital library.

Feasibility Issues

As we have seen, a suite of standards is emerging to govern the architecture of imaging systems. Moreover, a general understanding is growing about the nature and interrelation of system components in the architecture. Together, these factors generate confidence in our ability to achieve the vision of a large-scale preservation imaging system at Yale, in which material converted from microfilm becomes, in image format, much more accessible and valuable to the scholarly community.

Optimism about the overall feasibility of creating a viable system on the basis of a 10,000 volume image library, however, should not obscure the attendant risks. Not all aspects of the system components and their interrelationships are as well-understood as other parts; there are still many unknowns and uncertainties about the technology and its operation. By invoking a principle of design that we cited earlier—that the data created and stored are a library's principal asset in an automated system—and by asking how the data will be most at risk in this particular architecture, we can identify at least four areas of concern that might limit or impede the development of the system and so require extraordinary attention.

First, the area of principal concern is the integration of the mass storage devices. The technology of optical WORM storage is generally well-understood, but standard formats for the media have not yet emerged. In addition, experience in the industry is not yet very rich about the actual use of the technology and, in particular, about how to optimize the mix of optical and magnetic storage for particular applications and thereby to improve access time and to relieve contention for the drives in a multi-user environment.

On the face of it, the application that the library expects to develop seems special. Imaging applications have taken hold in office environments where the model document is a file folder, which contains a few items that one consults rapidly and moves on. The model document in the library, however, is the book of several hundred pages, over which one may pore for some time occasionally moving rapidly back and forth through the text to consult related points, citations, and so on. It seems reasonable to expect that the different patterns of document use in the library will stress the storage mechanisms in new and different ways that require solutions that may not yet be available. The question is: will experience bear out this hypothesis or not?

Second, the issue of periodically refreshing the data on optical storage media also remains a concern. Computer centers and computer vendors have now gone through many cycles of technical change in magnetic media, requiring data periodically to be recopied. The cycles of change in optical technology, however, have only begun to hit large production data centers. Because the technology is different and the volume of data stored on optical disk is orders of magnitude larger than on magnetic media, the familiar paradigms of data refreshment may not hold. Vendors will undoubtedly play a key role in promulgating appropriate and practical solutions, but, when asked, few are yet able to articulate such solutions.

The third major area of concern is the so-called document structure file. To the extent that the library indexes its image documents by page number or by the internal structure of the document, a major portion of its labor and investment is tied up in this structure file. The Office Document Architecture (ODA) and the SQL standards seem appropriate to govern its nature and organization and to make the data embedded in it transferrable to another system or implementation as necessary. However, it is already clear that this mode of indexing the content of images will require some revisions and extensions in the ODA standard. The investment in the image library and in the various indices incorporated in the document structure file will no doubt warrant the active participation of the Yale library not only in the specific implementation of the Office Document Architecture, but also in its extensions and development.

Finally, a fourth issue of potential concern emanates from the quality control process. The library community has expected for years that, given the right technology, microfilm will be easily convertible to digital form. The Yale project will surely generate much useful information about the technical ease (or difficulty) of the conversion process. But the quality of the product that results from the process will depend not only on the mechanics and technical wizardry of the conversion process itself, but also on the quality of the microfilm, on the quality needed for the digital images to serve as input for subsequent technical processes such as optical character recognition, and on the quality standards asserted by the library readers who will use the

System Architecture (continued)

digital images in their scholarly work. It will be important in this project, therefore, to capture and evaluate the ways in which preservation microfilm can or should be changed to generate a better digital image. Similarly, it will be important to find ways to evaluate the quality of the digitization process from the standpoint of related technical processes, such as character recognition, and it will be absolutely essential to evaluate the process from the standpoint of user satisfaction.

Plan of Work

The Yale digital preservation project will build on related work at the Library of Congress, the National Library of Medicine, Cornell University, and other library and archival institutions by focusing on the conversion of microfilm to digital imagery and by attempting to build a 10,000 volume digital library in the process. Much of the equipment and application software needed to construct a system at Yale is already available on the market from vendors or is in advanced stages of vendor development. The principal exception is the image presentation manager described in the previous section, which Yale has special expertise to develop and which will be generated as a specific product of the system work at Yale.

Of course, concerns about various ambiguous and uncertain aspects of the architecture of imaging systems highlight the risks of investing in relatively new and expensive technology and at the scale contemplated for the Yale project. However, those concerns also focus attention on the areas of opportunity that a project of Yale's scope and breadth offer for advancing our understanding of the potential and limits of the technology, for developing certain aspects of the technology and, overall, for measurably reducing the risks on investment for those libraries that follow on the imaging path. To mitigate the risks of the project and to enhance its prospects of yielding a valuable return, one solution is a deliberate and careful implementation plan with clear goals, measurable results and a phased approach in which all is not lost, or even necessarily at stake, if one encounters a problem or other obstacle.

Project Goals

The digital preservation project planned for the Yale University Library has three broad goals:

- Identify, develop and evaluate the means of creating and storing a digital library of preserved materials by converting images from microfilm to digitized form on a production scale of 10,000 volumes.
- Identify, develop and evaluate the means to provide, preserve and improve the distribution of, or physical access to, preserved materials in digital image form from both within and outside of the library.
- Identify, develop and evaluate the means to provide, preserve and improve intellectual access to library materials in digital image form.

Project Scope

The project will be divided into 6 phases over three years (see Figure 7). The first phase will extend from months 1—4 and will establish the overall organizational framework for the project. During the second phase, which will extend from months 5—8, the initial setup of hardware and software will occur. Phase three will establish a production level process for converting microfilm to digital images. It will begin in month 9 and end in month 18. The fourth phase of the

Plan of Work (continued)

project will also begin in month 9 but will end in month 21 when the digital image documents are made accessible to readers within the Yale library. The fifth phase will extend from months 22–33 and will establish access to the digital image documents from outside of the Yale library. The project will conclude in the sixth phase, during months 34–36, with completion of the 10,000 volume conversion and a final tuning and evaluation of all system components.

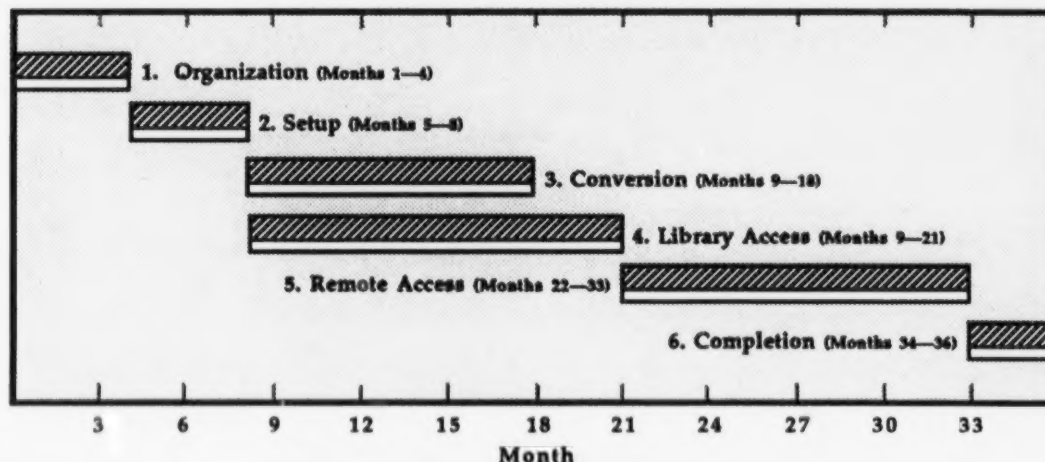


Figure 7. Project Phases

At each stage, a report will be generated to summarize the results to date and to evaluate the overall merit of the project. Because the digital image technology being investigated and demonstrated is still subject to rapid change and development, it is possible, though unlikely, that the results being sought here will emerge elsewhere or that the capabilities of the technology itself will overtake the assumptions on which the project is founded. Each report, therefore, will treat the end of phase as a milestone in which to ask and answer the question: is it worth continuing this project?

Phase 1: Organization

Objective: The objective of phase 1 is to create the overall organizational framework for the project and to prepare for work in subsequent phases.

Time Frame: Months 1–4.

Activities: During Phase 1, a Steering Committee will be created to provide general oversight and direction for the project. The committee will consist, at least initially, of the University Librarian, the Deputy University Librarian, the Associate University Librarians for Technical Services, Public Services and Collection Development, the Head of the Preservation Department, the University Director of Computing and Information Systems, the Director of Academic

Plan of Work (continued)

Computing, the Project Manager and four selected faculty members. The committee can adjust its membership over time as necessary.

The initial agenda of the committee will include the following items:

- establish the criteria for identifying which portions of the preservation microfilm collection will be converted to digital images.
- assist the Project Manager and members of the project team in establishing appropriate partnership relationships with vendors of document imaging equipment and applications.
- review the proposed budget for the project and assist in identifying and securing sources of funding.

The project manager will be the Head of the Systems Office in the library. The project team will consist initially of the Head of the Preservation Department in the library and the Associate Director for Systems, Technology and Planning in the Computing and Information Systems Department. During this phase, the project team will:

- identify the specific hardware and software that will be acquired for the project.
- establish appropriate partnership relationships with the vendors of the document imaging equipment and applications that will be acquired in the project.
- prepare a staffing plan for all subsequent phases of the project.
- prepare a budget for the entire project and work with the Steering Committee and other appropriate sources to obtain funding.

Prerequisites: This document, "From Microfilm to Digital Imagery," will provide the background and planning foundation for the initial work of the Steering Committee and the Project Team.

Products:

- A project budget.
- A funding plan.
- A report summarizing the results of work in this phase.

Phase 2: Setup

Objective: The objective of phase 2 is to begin setting up the working environment needed for the project.

Time Frame: Months 5-8.

Activities: The project team will:

- apply the criteria established by the Steering Committee and begin to select the portions of the microfilm collection to be converted.
- create and begin to implement a plan for evaluating whether the criteria used to select material for conversion creates a digital library that library clients find intellectually valuable.

Plan of Work (continued)

- hire needed staff.
- acquire, install and test the hardware and software for the conversion sub-system.
- acquire, install and test the high-speed digital printer and print server.
- acquire, install and test the document structure file component of the storage subsystem, and begin to acquire, install and test the other components of the storage subsystem.
- prepare the design and initial specifications for the image presentation server.

In addition, the team will contract with a service bureau to convert from microfilm to digital image form up to ten volumes that will be selected largely at random. The contract will establish a base line of costs against which to compare the costs of in-house processing. It will also provide experience in formulating and applying acceptance criteria for the in-house quality control process. Presumably, at this stage, the documents will be returned in image form without any indexing.

Prerequisites: In order for this phase to begin, there has to be an approved budget supplied with requisite funds to cover staff costs, the acquisition of hardware and software and the service bureau costs.

Products:

- 10 volumes will be available in digital form and ready for the indexing process within the conversion sub-system.
- A report summarizing the results of work in this phase, including an analysis of the service bureau experience and a plan to evaluate systematically throughout the duration of the project the merits of in-house conversion versus the use of a service bureau.

Phase 3: Conversion

Objective: The objective of phase 3 is to establish a production level process for converting microfilm to digital images.

Time Frame: Months 9-18.

Activities: The principal tasks for the project team in this phase include:

- develop procedures and train the operator(s) in the use of the digitizer and the quality control workstation, including its image enhancement features.
- select for digitization early in this phase a sample of microfilm reels that reflects variations in the quality of the microfilm process. Devise a method to use this sample both to exercise the quality control process and to begin to formulate conclusions about the ways that microfilming techniques affect the quality of the digitization process.
- devise and implement a quality control standard that accounts, in particular, for the future possibility of applying character

Plan of Work (continued)

recognition technology to the images generated in this conversion process.

- create an ongoing process for measuring user satisfaction with the quality control processes and standards.
- review the design of the page number index and the document structure index to insure that links can be created at some future time between these indices and supplementary contents notes that might be stored elsewhere in machine-readable form.
- develop procedures and train the operator(s) in the use of each of four different document indexing functions. The operator(s) should be able to (1) assign a unique identifier to the set of converted images that constitute a single document entity, (2) index the image document by page number, (3) index the image document by the internal structure of the document, and (4) create a bibliographic entity in the online catalog at Yale that describes the image document and identifies it in the digital image library by its unique identifier.
- devise methods of structuring and altering the indexing workflow to isolate the costs of each of the four forms of document indexing.
- develop procedures and train the operator(s) to commit the image document and the document indices to permanent file storage using, at this stage, a portable WORM drive.
- review criteria for selecting high use preservation materials for which a printed copy will be bound and returned to the shelf.
- develop procedures for and train the operator(s) to transfer the images of selected documents directly to the high quality image document printer. The preferred mode of transmission is via the high speed network.
- devise a method of structuring workflow to isolate the costs of digitization, quality control and printing, with no indexing.
- devise a method and an ongoing procedure for comparing the costs of in-house versus service bureau costs for digitization, quality control and indexing.
- index and commit to storage the 10 volumes digitized by the service bureau in phase 2.
- digitize, enhance, index and commit to storage 490 volumes.

Prerequisites: In order for this phase to begin, the conversion subsystem, the printer and print server and the document structure file component of the storage subsystem all have to be installed and tested.

Issues to be investigated: Work in this phase is designed to produce critical measures of quality and cost along several dimensions, including:

- the possible effects of microfilm quality on the quality of digitization.
- the implications of possible future character recognition for the quality control process in microfilm digitization.
- user satisfaction with the quality control processes and standards.
- the relative costs of in-house versus service bureau costs for digitization, quality control, and indexing.

Plan of Work (continued)

- the varying costs of the different levels of indexing.
- the costs of simply digitizing and printing without indexing.

It will be important to collect and monitor these various measures on an ongoing basis throughout the project.

Products:

- Operation of the conversion subsystem at production levels.
- 500 volumes digitized and fully indexed.
- A report summarizing the results of work in this phase, including an analysis of costs.

Phase 4: Library Access

Objective: The objective of phase 4 is to make the growing digital image library accessible to readers at workstations within the Yale library.

Time Frame: Months 9-21.

Activities: The principal tasks for the project team in this phase include:

- complete the acquisition, installation and testing of the storage subsystem.
- specify, acquire, install, and test the access workstations in the library.
- develop and test a prototype image presentation server capable of interacting with the access workstations in the library.
- design and implement a study to measure the value to library readers of each of the index tools available for the digital document compared to microfilm access.
- create and implement a plan for backing up data stored on magnetic media and for periodically refreshing data stored in optical and magnetic media.
- transfer all digitized volumes to the permanent storage system.
- devise a method for assessing the costs of adding the storage subsystem and of maintaining it on an ongoing basis.
- establish and implement methods and ongoing procedures for assessing the effects of access load on the performance of the image storage subsystem.
- devise and implement methods and ongoing procedures for tuning the performance of the image storage subsystem.
- establish and implement methods and ongoing procedures for monitoring and tuning the effects of image transmission on network performance.
- devise methods for assessing the costs of creating and maintaining the access subsystem.
- acquire, install and test the print requester.
- devise methods for assessing the costs of creating and maintaining the printing subsystem.
- after the completion of phase 2 and during months 16-18, digitize, enhance, index and commit to storage 1300 volumes.

Plan of Work (continued)

Prerequisites: In order for this phase to begin, the printer and print server and the document structure file component of the storage subsystem have to be installed and tested. The initial design and specifications for the image presentation server also have to be complete.

Issues to be investigated: Work in this phase is designed to produce critical measures of quality and cost along several dimensions, including:

- the incremental costs creating and maintaining the storage subsystem.
- the costs of backing up data stored on magnetic disk and of periodically refreshing data stored in optical and magnetic media.
- the incremental costs of creating and maintaining the access subsystem.
- the incremental costs of creating and maintaining the printing subsystem.
- the effects of access load on the storage subsystem.
- the effects of transmission load on the network.
- the value to readers in the library of the document indices.

It will be important to collect and monitor these various measures on an ongoing basis throughout the project.

Products:

- Image storage subsystem operating at production levels.
- Printing subsystem operating at production levels.
- Prototype image presentation manager.
- Access subsystem operating at production levels in the library.
- Print request server.
- 1300 volumes digitized and fully indexed.
- A report summarizing the results of work in this phase, including an analysis of costs.

Phase 5: Remote Access

Objective: The objective of phase 5 is to establish access to the digital image documents from outside of the Yale library.

Time Frame: Months 22-33.

Activities: The principal tasks for the project team in this phase include:

- extend the operation of the image presentation server to be capable of interacting with selected types of workstations outside the library.
- make the imaging access subsystem available through the image presentation server to selected types of workstations that are connected to the campus network.

Plan of Work (continued)

- devise methods for assessing the startup and ongoing costs of adding access to the image system from outside the library.
- design and implement a study to measure the value to users outside of the library of each of the index tools available for the digital document.
- digitize, enhance, index and commit to storage 7800 volumes.

Prerequisites: In order for this phase to begin, the prototype version of the image presentation server has to be complete and operational.

Issues to be investigated: Work in this phase is designed to produce critical measures of the value to readers outside of the library of the document indices. It will be important to collect and monitor these measures on an ongoing basis during this and the last phase of the project.

Products:

- Production level image presentation manager.
- Operation of the access subsystem on workstations outside the library.
- 7800 volumes digitized and fully indexed.
- A report summarizing the results of work in this phase, including an analysis of costs.

Phase 6: Completion

Objective: The objective of phase 6 is to complete the 10,000 volume conversion and perform a final tuning and evaluation of all system components.

Time Frame: Months 34-36.

Activities: The principal tasks for the project team in this phase include:

- review and assess all measures of quality and performance in each of the subsystems.
- digitize, enhance, index and commit to storage 400 volumes.

Prerequisites: In order for this phase to begin, the production version of the image presentation server has to be complete and operational.

Products:

- A final report.
- 400 volumes digitized and fully indexed.

Conclusion

The Yale University Library envisions a future in which digital image technology comprises a critical tool in the process of preserving access to the deteriorating materials in its rich and valuable collections. Just at a time when the deterioration of documents composed of acidic paper is accelerating at an alarming rate at Yale and at other major research libraries across the country, the architecture of digital image systems is becoming well understood and established, and the system components are increasingly well-integrated and readily available. The Library of Congress, the National Library of Medicine, Cornell University, and other library and archival institutions have in various ways made major contributions to the development of imaging technology and in applying it to preserve and improve general access to the national and international store of recorded knowledge. The time is right for another major step toward making digital imagery a practical library tool.

The Yale digital preservation project will build on related work elsewhere, particularly at Cornell University, by focusing on the conversion of microfilm to digital imagery and by attempting to build a 10,000 volume digital library in the process. Much of the equipment and application software needed to construct the imaging system at Yale is available or in final stages of vendor development. However, the Yale preservation imaging system will be both capital intensive and, particularly in the indexing phases of the conversion process, labor intensive. Its success will require major support from funding agencies and significant and productive partnerships with members of the vendor community. The outcome for all participants in the project will be a substantial body of information about the costs and benefits of digital imagery. Libraries, and the vendors who work with them, will depend on such information in the future to make prudent, practical decisions about incorporating imaging technology into the daily routine of library work and of scholarly access to library materials.

Notes

Acknowledgements: In preparing this planning report, I received much help from many people in many private and university institutions. They are too numerous to name; I am grateful to them all. However, I do want to acknowledge my colleagues in the Yale University Library Systems Office, Merri Beth Lavagnino and Greg Kaisen, for their considerable help. Also I thank the Commission for Preservation and Access for its support in contracting the Yale University Library to prepare this report.

- 1 Nancy E. Gwinn, ed., *Preservation Microfilming: A Guide for Librarians and Archivists* (Chicago: American Library Association, 1987), p. xxvi. John C. Mallinson, "On the Preservation of Human- and Machine-Readable Records," *Information Technology and Libraries*, 7:1 (March 1988): 22.
- 2 Michael Lesk, "Digital Imagery, Preservation and Access," *Information Technology and Libraries*, 9:4 (December 1990): 307.
- 3 *Ibid.*, 308.
- 4 Constance C. Gould, *Information Needs in the Humanities: An Assessment*, (Mountain View, Ca.: Research Libraries Group), p. 51. Constance C. Gould and Mark Handler, *Information Needs in the Social Sciences: An Assessment*, (Mountain View, Ca.: Research Libraries Group, 1989). Constance C. Gould, *Information Needs in the Sciences: An Assessment*, (Mountain View, Ca.: Research Libraries Group, 1991).
- 5 Eric Almquist, "An Examination of Work-Related Information Acquisition and Usage among Scientific, Technical and Medical Fields," results of a study commissioned by the Faxon Institute for Advanced Studies in Scholarly and Scientific Communications, presented at the 1991 Faxon Institute Annual Conference, *Creating User Pathways to Electronic Information*, April 1991, Reston, Virginia, Exhibit 2.
- 6 *Ibid.*, Exhibits 9, 22 and 25.
- 7 See, for example, F. W. Lancaster, *The Measurement and Evaluation of Library Services* (Washington, D.C.: Information Resources Press, 1977), p. 313. Lancaster refers to a series of studies which support the argument that "selection of an information source is based almost exclusively on accessibility, the most accessible source (channel) being chosen first; considerations of quality and reliability are secondary, although these factors are important in influencing the degree to which the user is willing to accept information supplied from a particular source."
- 8 Lawrence Dowler, "Conference on Research Trends and Library Resources," *Harvard Library Bulletin*, 1:2 (Summer, 1990): 8.
- 9 Lesk, *op. cit.*, 303.

Plan of Work (continued)

- 10 Blipping, the placement of special markings on selected frames, can facilitate the use of microfilm, especially if the markings are coded and indexed in a system for computer-aided retrieval. The technique, however, has been applied primarily to 16 mm film. UMI reportedly has defined blipping guidelines for its internal use of 35 mm film. Whether the guidelines will become widely available and applied more generally to the use of 35 mm preservation microfilm in libraries remains to be seen.
- 11 Character recognition from digitally scanned images raises a challenging set of issues. First, although recognition algorithms are constantly improving, the accuracy of the current technology has been disappointing. Second, much work still needs to be done to improve the ability of character recognition algorithms to deal with the wide variety of typography and languages that one encounters in library preservation materials. Third, given successful recognition of the alphanumeric characters which compose a text image, one needs for purposes of storage and retrieval to relate the characters to the structure and layout of the original document, including any pictures, drawings, diagrams or other graphic images that may be embedded in it.

How these issues are ultimately addressed and resolved will greatly affect the accessibility of documents stored in electronic form. We intend to be mindful of and informed about the implications of the decisions and choices we make in the current project for the future application of character recognition technology. Indeed, we intend to take every reasonable precaution to insure the success of such an application to the digital images we produce. However, given the complexity and relative immaturity of character recognition technology, we regard its application as related to but essentially out of scope of the present project.

For a useful discussion of some of the issues associated with applying character recognition to library materials, see Stuart Weibel, John Handley and Charles Huff, "Automated Document Architecture Processing and Tagging," in Donald L. Blamberg, Carol L. Dowling, and Claudia V. Weston, eds., *Proceedings of the Conference on Application of Scanning Methodologies in Libraries*, (Beltsville, Maryland: National Agricultural Library, 1989), pp. 3-14.

- 12 See, for example, Karen Markey, *Subject Searching in Library Catalogs: Before and after the Introduction of Online Catalogs* (Dublin, Ohio: Online Computer Library Center, 1984), pp. 75-117; Richard Van Orden, "Context-Enriched Access to Electronic Information: Summaries of Selected Research," *Library Hi Tech* 8:3 (1990): 27-32; and Martin Dillon and Patrick Wenzel, "Retrieval Effectiveness of Enhanced Bibliographic Records," *Library Hi Tech* 8:3 (1990): 43-46.

Plan of Work (continued)

- 13 On the continued usefulness of microfilm in industry, see Whitney S. Minkler, "Optical Disks vs. Micrographics. Is it an Adversarial Problem," *Micrographics and Optical Technology*, 7:4 (1989): 141-149; and John Blake, "War over Optical Disk or Microfilm Ends, Future Contains Multiple Media," *Micrographics and Optical Technology*, 8:3 (1990): 141-143.
- 14 See, for example, Clifford A. Lynch and Edwin B. Brownrigg, "Library Applications of Electronic Imaging Technology," *Information Technology and Libraries*, 5:2 (June 1986): 100-102.
- 15 Lesk, *op. cit.*, 307.
- 16 Felix P. Krayeski, "Transition of an Image System: From Paper to Microfiche to Optical Disk," Congressional Research Service, The Library of Congress, June 1990. Frank L. Walker and George R. Thoma, "Access Techniques for Document Image Databases," *Library Trends*, 38:4 (Spring 1990): 751-786. Frank L. Walker, "Issues in Document Conversion," in Donald L. Blamberg, Carol L. Dowling, and Claudia V. Weston, eds., *Proceedings of the Conference on Application of Scanning Methodologies in Libraries*, (Beltsville, Maryland: National Agricultural Library, 1989), pp. 45-60.
- 17 For a useful overview of relevant standards, see Gerry Walter, "Standards Help Advance Document Management System Progress," *Optical Memory News*, (May, 1990): 25-27.
- 18 Presenting a document structure index in a standard relational form will eventually help facilitate the creation of links between the digital image document and any content information stored in other machine-readable sources, such as the online bibliographic description of the document.
- 19 Lesk, *op. cit.*, 306.